DS-GA 1013: Project
Intelligible Principal Component Analysis

Kho, Lee *ltk224*
Peeters, Jonas *jp5642*

March 1, 2020

We want to research different methods of retrieving intelligible principal components when looking at gene expressions. We consider the sparse principal component analysis model introduced by Zou, Hastie and Tibshirani (2006) the key piece of literature for this paper. After reading our paper, the reader should be informed about the structural differences between SPCA approaches, its potential applications, and benchmarks against other methods. As an application we will perform SPCA on UC Irvine's *gene expression cancer RNA-seq data set* (Dua & Graff, 2017) and interpret its results. We have the following hypotheses we want to check.

Hypothesis 1 *SPCA is the primus inter pares of intelligible PCA methods.*[1]

Hypothesis 2 *SPCA is able to condense UC Irvine's gene expression cancer RNA-seq data set to interpretable principal components revealing gene groupings.*

Context

Jolliffe (1986) introduced the ubiquitous dimension-reduction technique PCA in 1986. It has since been a staple in statistics curricula and academic research. One of the most relevant drawbacks to using principal component analysis (henceforth PCA) is the difficulty of interpreting separate principal components as they are a linear combination of all the original variables. Since its creation, several attempts have been made at making the principal components more intelligible. One of those is the rotation of some or all the components to simplify them. However, this approach is only permissible on a certain subset of components, namely those with nearly equal variance (Jolliffe, 1989, 1995).

Another approach frequently applied in order to increase the readability of the components is simply setting the loadings of variables with a small-magnitude to zero. This method could lead to misleading results for a series of reasons (Cadima & Jolliffe, 1995). Vines (2000) devised a method that limits the individual loadings to -1, 0 or 1. Jolliffe, Trendafilov, and Uddin (2003) introduced SCoTLASS, which is a modified PCA technique that is based on the least absolute shrinkage and selection operator (or lasso). Hence, penalizing small loadings on many different variables.

One of the later developments is sparse PCA (henceforth SPCA) developed by Zou, Hastie and Tibshirani (2006). Similarly to SCoTLASS this method incorporates the lasso method, but in addition, they reformulate the PCA in a regression-type optimization problem. This method achieves overall sparse loadings with limited variable overlap between the different principal components. In the years following its first publication, SPCA has been improved upon in a plethora of ways (Shen & Huang, 2008; Journée, Nesterov, Richtárik, & Sepulchre, 2010).

Data set and methodology

The UCI's dataset contains a random extraction of gene expressions of patients with different types of tumors, such as breast, kidney, cancer, colon, lung and prostate cancer. The data set has 801 observations and 20531 features which makes it interesting as an SPCA test case. We will use SPCA to group different gene expressions and their effect on the development of different types of cancer. Besides SPCA, we will also apply the SCoTLASS, thresholding of loadings and normal PCA methods and compare their principal components.

[1] We will be looking specifically at SPCA, SCoTLASS, thresholding of loadings and normal PCA.

References

Cadima, J., & Jolliffe, I. T. (1995, jan). Loading and correlations in the interpretation of principle compenents. *Journal of Applied Statistics*, *22*(2), 203–214. doi: 10.1080/757584614

Dua, D., & Graff, C. (2017). *UCI machine learning repository.* Retrieved from `http://archive.ics.uci.edu/ml`

Jolliffe, I. T. (1986). Principal components in regression analysis. In *Principal component analysis* (pp. 129–155). New York, NY: Springer New York. Retrieved from `https://doi.org/10.1007/978-1-4757-1904-8_8` doi: 10.1007/978-1-4757-1904-8_8

Jolliffe, I. T. (1989). Rotation of iii-defined principal components. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *38*(1), 139–147.

Jolliffe, I. T. (1995). Rotation of principal components: choice of normalization constraints. *Journal of Applied Statistics*, *22*(1), 29-35. Retrieved from `https://doi.org/10.1080/757584395` doi: 10.1080/757584395

Jolliffe, I. T., Trendafilov, N. T., & Uddin, M. (2003, sep). A modified principal component technique based on the LASSO. *Journal of Computational and Graphical Statistics*, *12*(3), 531–547. doi: 10.1198/1061860032148

Journée, M., Nesterov, Y., Richtárik, P., & Sepulchre, R. (2010). Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research*, *11*(Feb), 517–553.

Shen, H., & Huang, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of multivariate analysis*, *99*(6), 1015–1034.

Vines, S. K. (2000, jan). Simple principal components. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *49*(4), 441–451. doi: 10.1111/1467-9876.00204

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, *15*(2), 265-286. Retrieved from `https://doi.org/10.1198/106186006X113430` doi: 10.1198/106186006X113430